Corpus Linguistics: 
A Practical Introduction

Nadja Nesselhauf, October 2005 (last updated September 2011)

1) Corpus Linguistics and Corpora
- What is corpus linguistics (I)?
- What data do linguists use to investigate linguistic phenomena?
- What is a corpus?
- What is corpus linguistics (II)?
- What corpora are there?
- What corpora are available to students of English at the University of Heidelberg?
(For a list of corpora available at the Department of English click here)

2) Corpus Software
- What software is there to perform linguistic analyses on the basis of corpora?
- What can the software do?
- A brief introduction to an online search facility (BNC)
- A step-to-step introduction to WordSmith Tools

3) Exercises (I and II)
- I Using the WordList function of WordSmith
- II Using the Concord function of WordSmith

4) How to conduct linguistic analyses on the basis of corpora: two examples
- Example 1: Australian English vocabulary
- Example 2: Present perfect and simple past in British and American English
- What you have to take into account when performing a corpuslinguistic analysis

5) Exercises (III)
- Exercise III.1
- Exercise III.2

6) Where to find further information on corpus linguistics
1) Corpus Linguistics and Corpora

What is corpus linguistics (I)?
Corpus linguistics is a method of carrying out linguistic analyses. As it can be used for the investigation of many kinds of linguistic questions and as it has been shown to have the potential to yield highly interesting, fundamental, and often surprising new insights about language, it has become one of the most wide-spread methods of linguistic investigation in recent years.

What data do linguists use to investigate linguistic phenomena?
Roughly, four types of data for linguistic analysis can be distinguished:
1) data gained by intuition
   a) the researcher’s own intuition (“introspection”)
   b) other people’s (“informant’s”) intuition (accessed, for example, by elicitation tests)
2) naturally occurring language
   a) randomly collected texts or occurrences (“anecdotal evidence”)
   b) systematic collections of texts (“corpora”)
(For further reading on corpora vs. intuition, see Fillmore 1992)

What is a corpus?
A corpus can be defined as a systematic collection of naturally occurring texts (of both written and spoken language).
   “Systematic” means that the structure and contents of the corpus follows certain extralinguistic principles (“sampling principles”, i.e. principles on the basis of which the texts included were chosen). For example, a corpus is often restricted to certain text types, to one or several varieties of English, and to a certain time span. If several subcategories (e.g. several text types, varieties etc.) are represented in a corpus, these are often represented by the same amount of text. “Systematic” also means that information on the exact composition of the corpus is available to the researcher (including the number of words in each category and in the whole corpus, how the texts included in the corpus were sampled etc).
   Although “corpus” can refer to any systematic text collection, it is commonly used in a narrower sense today, and is often only used to refer to systematic text collections that have been computerized.

What is corpus linguistics (II)?
Corpus linguistics thus is the analysis of naturally occurring language on the basis of computerized corpora. Usually, the analysis is performed with the help of the computer, i.e. with specialised software, and takes into account the frequency of the phenomena investigated.

What corpora are there?
There are many types of corpora, which can be used for different kinds of analyses (cf. Kennedy 1998). Some (not necessarily mutually exclusive) examples of corpus types are (for a description of the individual corpora see below):

- general/reference corpora (vs. specialized corpora)
  (e.g. BNC = British National Corpus, or Bank of English):
    aim at representing a language or variety as a whole (contain both spoken and written language, different text types etc.)
- historical corpora (vs. corpora of present-day language)
  (e.g. Helsinki Corpus, ARCHER)
aim at representing an earlier stage or earlier stages of a language
- regional corpora (vs. corpora containing more than one variety)
  (e.g. WCNZE = Wellington Corpus of Written New Zealand English)
  aim at representing one regional variety of a language
- learner corpora (vs. native speaker corpora)
  (e.g. ICLE = International Corpus of Learner English)
  aim at representing the language as produced by learners of this language
- multilingual corpora (vs. one-language corpora)
  aim at representing several, at least two, different languages, often with the same text
  types (for contrastive analyses)
- spoken (vs. written vs. mixed corpora)
  (e.g. LLC = London-Lund Corpus of Spoken English)
  aim at representing spoken language

A further distinction of corpus types refers not to the texts that have been included in the
  corpus, but to the way in which these texts have been treated:
- annotated corpora (vs. orthographic copora)
  in annotated corpora, some kind of linguistic analysis has already been performed on
  the texts, such as sentence analysis, or, more commonly, word class classification

What corpora are available to students of English at the University of Heidelberg?

1) Generally accessible corpora:

Two large general corpora of English are accessible to everyone via the World Wide Web. These are the Collins Wordbanks Online English corpus and the British National Corpus.

a) The Collins Wordbanks Online English corpus contains 56 million words of
  contemporary written and spoken text, both British and American English, of a variety
  of text types. Of these, 36 million are British written texts, 10 million American
  written texts and 10 million American spoken texts. The user can select either one or
  two or all three subcorpora for the analysis. The corpus is accessible at:
  http://www.collins.co.uk/Corpus/CorpusSearch.aspx.

b) The British National Corpus (BNC) contains 100 million words of contemporary
  British English, of which 90 million are written and 10 million spoken texts (of a
  variety of different text types). For simple searches, the corpus is accessible at:
  http://sara.natcorp.ox.ac.uk/lookup.html (For an explanation of the search facility, see
  below. Much more complex searches are possible on the basis of the CD-Rom version
  of the corpus, also available at the Department).

A few smaller and more specialised corpora have also been made available on the internet. An example is MICASE:

c) MICASE (The Michigan Corpus of Academic Spoken English) contains 1.8 million
  words of (transcribed) academic speech, recorded at the University of Michigan
  between 1997 and 2001. The corpus contains different academic speech events, and
  both native and non-native speaker language. It is available at:
  http://micase.umdl.umich.edu/m/micase/.

2) Corpora at the English Department
A great number of corpora are available for students of English, the most important of which are listed below. Many of these are accessible on one of the two “corpus computers” located in the seminar library (“library” in the table). The corpus computers are located in the second room, in the right-hand corner, just before the linguistic section of the library. Other corpora are available at the Lehrstuhl Prof. Busse.

If you intend to carry out research with one of the corpora, please contact Nadja Nesselhauf:

Nadja.Nesselhauf@urz.uni-heidelberg.de

For office hours, see: http://www.as.uni-heidelberg.de/personen/

You will then receive a password for the corpus computers or receive access to one of the corpora available at the Lehrstuhl.

List of corpora available at the Department of English

Most of the available corpora are stored directly in C: on the corpus computers (the path for ICE New Zealand, for example, simply is C:\icenz), many are stored in C:\ICAME\texts. Many manuals are also stored on the corpus computer, in the same directory as the corpus in question.

Extensive lists of existing corpora can be found on Michael Barlow’s “Corpus Linguistics Site” at http://www.athel.com/corpus.html and on the “English language corpora and corpus resources” page at http://www.natcorp.ox.ac.uk/corpora.html.
<table>
<thead>
<tr>
<th>Corpus</th>
<th>Varieties</th>
<th>Time (span)</th>
<th>Number of words</th>
<th>Text type(s)</th>
<th>Available</th>
<th>Further information available at / in:</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Large general corpora:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BNC (British National Corpus)</td>
<td>BrE</td>
<td>PDE</td>
<td>100 m</td>
<td>90% written, 10% spoken, many diff. text types</td>
<td>General access</td>
<td>simple searches: <a href="http://www.natcorp.ox.ac.uk/using/index.xml?ID=simple">http://www.natcorp.ox.ac.uk/using/index.xml?ID=simple</a></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>more advanced searches: <a href="http://bncweb.lancs.ac.uk/bncwebSignup/userLogin.php">http://bncweb.lancs.ac.uk/bncwebSignup/userLogin.php</a></td>
</tr>
<tr>
<td>COCA (The Corpus of</td>
<td>AmE</td>
<td>1990-2011</td>
<td>425 m</td>
<td>Various spoken and written text types</td>
<td>General access</td>
<td><a href="http://www.americancorpus.org/">http://www.americancorpus.org/</a></td>
</tr>
<tr>
<td>Contemporary American English</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ANC (American National Corpus)</td>
<td>AmE</td>
<td>PDE</td>
<td>20 m</td>
<td>Written and spoken</td>
<td>librarian</td>
<td><a href="http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2003T20">http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2003T20</a></td>
</tr>
<tr>
<td><strong>Comparable corpora of written British and American English:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LOB</td>
<td>BrE</td>
<td>1961</td>
<td>1 m</td>
<td>2,000-word samples from diff. written text types</td>
<td>Corpus computer (library)</td>
<td><a href="http://khnt.hit.uib.no/icame/manuals/index.htm">http://khnt.hit.uib.no/icame/manuals/index.htm</a> manual in library next to corpus computers</td>
</tr>
<tr>
<td>Brown</td>
<td>AmE</td>
<td>1961</td>
<td>1 m</td>
<td>same as LOB</td>
<td>Corpus computer (library)</td>
<td><a href="http://khnt.hit.uib.no/icame/manuals/index.htm">http://khnt.hit.uib.no/icame/manuals/index.htm</a> manual in library next to corpus computers</td>
</tr>
<tr>
<td>FLOB</td>
<td>BrE</td>
<td>1991</td>
<td>1 m</td>
<td>same as LOB</td>
<td>Corpus computer (library)</td>
<td><a href="http://khnt.hit.uib.no/icame/manuals/index.htm">http://khnt.hit.uib.no/icame/manuals/index.htm</a> manual in library next to corpus computers</td>
</tr>
<tr>
<td>Frown</td>
<td>AmE</td>
<td>1992</td>
<td>1 m</td>
<td>same as LOB</td>
<td>Corpus computer (library)</td>
<td><a href="http://khnt.hit.uib.no/icame/manuals/index.htm">http://khnt.hit.uib.no/icame/manuals/index.htm</a> manual in library next to corpus computers</td>
</tr>
<tr>
<td><strong>Corpora of other varieties, comparable to the above in composition:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WWC (Wellington Corpus of</td>
<td>NZE</td>
<td>1986-1990</td>
<td>1 m</td>
<td>same as LOB</td>
<td>corpus computer (library)</td>
<td><a href="http://khnt.hit.uib.no/icame/manuals/index.htm">http://khnt.hit.uib.no/icame/manuals/index.htm</a> manual in library next to corpus computers</td>
</tr>
<tr>
<td>Written New Zealand English)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Region</td>
<td>Language</td>
<td>Start-End Year</td>
<td>Size</td>
<td>Description</td>
<td>Access Information</td>
<td></td>
</tr>
<tr>
<td>-----------------------------</td>
<td>----------</td>
<td>----------------</td>
<td>------</td>
<td>------------------------------------------------------------------</td>
<td>-------------------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>Kolhapur</td>
<td>IndE</td>
<td>1978</td>
<td>1 m</td>
<td>same as LOB</td>
<td>corpus computer (library)</td>
<td></td>
</tr>
<tr>
<td>ACE (The Australian Corpus of English)</td>
<td>AuE</td>
<td>1986</td>
<td>1 m</td>
<td>same as LOB</td>
<td>corpus computer (library)</td>
<td></td>
</tr>
<tr>
<td><strong>Other regional corpora:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ICE-New Zealand</td>
<td>NZE</td>
<td>1989-1994</td>
<td>1 m</td>
<td>diff. spoken (600,000) and written (400,000) genres</td>
<td>corpus computer (library)</td>
<td></td>
</tr>
<tr>
<td>ICE-Singapore</td>
<td>SingE</td>
<td></td>
<td>1 m</td>
<td>diff. spoken (600,000) and written (400,000) genres</td>
<td>General access</td>
<td></td>
</tr>
<tr>
<td>ICE-India</td>
<td>IndE</td>
<td></td>
<td>1 m</td>
<td>diff. spoken (600,000) and written (400,000) genres</td>
<td>General access</td>
<td></td>
</tr>
<tr>
<td>ICE-Philippines</td>
<td>PhiE</td>
<td></td>
<td>1 m</td>
<td>diff. spoken (600,000) and written (400,000) genres</td>
<td>General access</td>
<td></td>
</tr>
<tr>
<td>ICE-GB (Great Britain)</td>
<td>BrE</td>
<td>1990-1993</td>
<td>1 m</td>
<td>diff. spoken (600,000) and written (400,000) genres</td>
<td>General access</td>
<td></td>
</tr>
<tr>
<td>ICE-Canada</td>
<td>CanE</td>
<td></td>
<td>1 m</td>
<td>diff. spoken (600,000) and written (400,000) genres</td>
<td>General access</td>
<td></td>
</tr>
<tr>
<td>ICE-Hong Kong</td>
<td>HKE</td>
<td></td>
<td>1 m</td>
<td>diff. spoken (600,000) and written (400,000) genres</td>
<td>General access</td>
<td></td>
</tr>
<tr>
<td>ICE-Ireland</td>
<td>IrE</td>
<td></td>
<td>1 m</td>
<td>diff. spoken (600,000) and written (400,000) genres</td>
<td>General access</td>
<td></td>
</tr>
<tr>
<td>ICE-East Africa</td>
<td>Kenya n &amp; Tanz. Engl.</td>
<td>1990-1996</td>
<td>1.3 m</td>
<td>400,000 w. written Kenyan English, 400,000 w. written Tanzanian English, 500,000 words spoken (Kenyan &amp; Tanzanian)</td>
<td>Generally available</td>
<td></td>
</tr>
</tbody>
</table>

*ICE-GB (Great Britain)*

http://www.ucl.ac.uk/english-usage/ice/index.htm

*ICE-Canada*

http://ice-corpora.net/ice/index.htm

*ICE-Hong Kong*

http://ice-corpora.net/ice/index.htm

*ICE-Ireland*

http://ice-corpora.net/ice/index.htm

*ICE-East Africa*

http://ice-corpora.net/ice/index.htm
<table>
<thead>
<tr>
<th>Corpus Name</th>
<th>Language</th>
<th>Date</th>
<th>Size</th>
<th>Description</th>
<th>Access/Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>WSC (Wellington Corpus of Spoken New Zealand English)</td>
<td>NZE</td>
<td>1988-1994</td>
<td>1 m</td>
<td>2,000 word extracts of formal, semi-formal, and informal speech</td>
<td><a href="http://khnt.hit.uib.no/icame/manuals/index.htm">http://khnt.hit.uib.no/icame/manuals/index.htm</a></td>
</tr>
<tr>
<td>Spoken corpora:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LLC (London-Lund Corpus of Spoken English)</td>
<td>BrE</td>
<td>1975-1988</td>
<td>500,000</td>
<td>transcribed spoken lg., various text types</td>
<td><a href="http://khnt.hit.uib.no/icame/manuals/index.htm">http://khnt.hit.uib.no/icame/manuals/index.htm</a></td>
</tr>
<tr>
<td>Santa Barbara Corpus of Spoken American English</td>
<td>AmE</td>
<td>PDE</td>
<td>300,000</td>
<td>Transcribed spoken interaction, from all over the U.S.</td>
<td><a href="http://www.linguistics.ucsb.edu/research/sbcorpus.html">http://www.linguistics.ucsb.edu/research/sbcorpus.html</a></td>
</tr>
<tr>
<td>MICASE (The Michigan Corpus of Academic Spoken English)</td>
<td>AmE</td>
<td>1997-2001</td>
<td>1.8 m</td>
<td>Transcribed academic speech (native and non-native)</td>
<td><a href="http://micase.elicorpora.info/">http://micase.elicorpora.info/</a> or <a href="http://quod.lib.umich.edu/m/micase//">http://quod.lib.umich.edu/m/micase//</a></td>
</tr>
<tr>
<td>Historical corpora:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Helsinki Corpus (diachronic part)</td>
<td>BrE</td>
<td>750-1700</td>
<td>1.5 m</td>
<td>different text types</td>
<td><a href="http://khnt.hit.uib.no/icame/manuals/index.htm">http://khnt.hit.uib.no/icame/manuals/index.htm</a></td>
</tr>
<tr>
<td>Helsinki Corpus of Older Scots</td>
<td>Scots</td>
<td>1450-1700</td>
<td>870,000</td>
<td>several text types</td>
<td><a href="http://khnt.hit.uib.no/icame/manuals/index.htm">http://khnt.hit.uib.no/icame/manuals/index.htm</a></td>
</tr>
<tr>
<td>ARCHER (A Representative Corpus of Historical English Registers)</td>
<td>BrE &amp; AmE</td>
<td>1650-1990</td>
<td>1.7 m</td>
<td>several diff. text types, both written and speech-based</td>
<td><a href="http://www.llc.manchester.ac.uk/research/projcts/archer/archer3_1/">http://www.llc.manchester.ac.uk/research/projcts/archer/archer3_1/</a></td>
</tr>
<tr>
<td>Corpus Name</td>
<td>Language</td>
<td>Date Range</td>
<td>Number of Items</td>
<td>Text Types</td>
<td>Access Method</td>
</tr>
<tr>
<td>-------------------------------------------------</td>
<td>----------</td>
<td>-----------------------</td>
<td>-----------------</td>
<td>-------------------------------------------------</td>
<td>---------------------------</td>
</tr>
<tr>
<td>CEECS (Corpus of Early English Correspondence Sampler)</td>
<td>BrE</td>
<td>1418-1680</td>
<td>450,000</td>
<td>letters</td>
<td>corpus computer (library)</td>
</tr>
<tr>
<td>PCEEC (Parsed Corpus of Early English Correspondence Sampler)</td>
<td>BrE</td>
<td>1418-1680</td>
<td>2.2 m</td>
<td>Letters</td>
<td>General access</td>
</tr>
<tr>
<td>DCPSE</td>
<td>BrE</td>
<td>1950s-1990s</td>
<td>800,000</td>
<td>Various spoken text types</td>
<td>librarian</td>
</tr>
<tr>
<td>CED</td>
<td>BrE</td>
<td>1560-1760</td>
<td>1.2 m</td>
<td>various speech-based text types</td>
<td>contact Prof. Busse</td>
</tr>
<tr>
<td>Lampeter Corpus of Early Modern English Tracts</td>
<td>BrE</td>
<td>1640-1740</td>
<td>1.1 m</td>
<td>tracts from various areas (religion, law, science etc.)</td>
<td>corpus computer (library)</td>
</tr>
<tr>
<td>ICAMET (Innsbruck Computer Archive of Machine-Readable English Texts)</td>
<td>BrE</td>
<td>1100-1688</td>
<td>5.6 m</td>
<td>prose 1100-1500, letters 1386-1688</td>
<td>corpus computer (library)</td>
</tr>
<tr>
<td>COHA (The Corpus of Historical American English)</td>
<td>AmE</td>
<td>1810-2009</td>
<td>400 m</td>
<td>Various text types</td>
<td>General access</td>
</tr>
<tr>
<td>Concer</td>
<td>BrE</td>
<td>19th c.</td>
<td>Various text types</td>
<td></td>
<td>Contact Prof. Busse</td>
</tr>
<tr>
<td>18th Century Fiction</td>
<td>BrE</td>
<td>18th c.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19th Century Fiction</td>
<td>BrE</td>
<td>19th c.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scientific and Medical Writings in Old and Middle English</td>
<td>BrE</td>
<td>OE and ME</td>
<td></td>
<td>Scientific and medical writing</td>
<td>librarian</td>
</tr>
<tr>
<td>Corpus of child language:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td><strong>POW (Polytechnic of Wales Corpus)</strong></td>
<td>BrE</td>
<td>1978-1984</td>
<td>65,000</td>
<td>speech of children between 6 and 12 (spontaneous and interviews)</td>
<td>corpus computer (library)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Learner corpus:</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ICLE (International Corpus of Learner English)</strong></td>
<td>learner Engl.</td>
<td>1990s</td>
<td>2.5 m</td>
<td>argumentative essays written by learners with different L1s</td>
<td>corpus computer (library)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Early Modern English Medical Texts</th>
<th>BrE</th>
<th>EMO</th>
<th>2 m</th>
<th>Different types of medical writing</th>
<th>librarian</th>
<th><a href="http://www.helsinki.fi/varieng/CoRD/corpora/CEEM/MEMTindex.html">http://www.helsinki.fi/varieng/CoRD/corpora/CEEM/MEMTindex.html</a></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Middle English Medical Texts</strong></td>
<td>BrE</td>
<td>ME</td>
<td>500,000</td>
<td>Medical writing (surgical treatises, remedy books, recipes etc.)</td>
<td>librarian</td>
<td><a href="http://www.helsinki.fi/varieng/CoRD/corpora/CEEM/MEMTindex.html">http://www.helsinki.fi/varieng/CoRD/corpora/CEEM/MEMTindex.html</a></td>
</tr>
</tbody>
</table>
2) Corpus Software

What software is there to perform linguistic analyses on the basis of corpora?

Two types of software for corpus analysis can be distinguished in principle: software that is tailored to one specific corpus, and software that can be used with almost any kind of corpus. Examples of the former type are two software programs that have been tailored to the British National Corpus, namely SARA and BNCWeb (accessible on the left corpus computer in the seminar library of the English Department at the University of Heidelberg). A further example is ICE-CUP, which has been tailored to ICE-GB. A subgroup of this type is software which allows searches in one specific corpus over the internet (such as the online search facilities provided for the BNC and the Collins Wordbanks Online English mentioned above). Examples of the second type of more generally usable software are MonoConc Pro (demo available at http://www.camsoftpartners.co.uk/monoconc.htm) and WordSmith Tools, which is probably the most widely used corpus software. WordSmith is available for corpus research on the right corpus computer in the library; its use will be explained in detail below.

What can the software do?

While there are many differences between the software packages designed for corpus analysis, certain basic functions can be performed by practically all the available software. For most kinds of linguistic analyses, the most important one of these is the possibility of searching the corpus in question for the occurrence of certain strings (i.e. words or phrases). As output, the software then usually gives information on the number of these strings occurring in the corpus, on the part of the corpus and/or text in which they were found, and so-called concordance-lines, which show the string in question in context (with the search term(s) highlighted), such as in this example (the result of a WordSmith query for the string stark in ACE, the Australian Corpus of English):

![Concordance Example](image)

The number on the left indicates the total number of occurrences (9), on the right, further information such as in which file the instances were found is provided (what type of text the...
different file names refer to can be found out by consulting the manual; ace_j.txt, for example, means that the instance occurs in a science text, ace_k.txt that it occurs in a fiction text etc.). Two further basic functions that can be performed by almost all corpus software are sorting (for example according to the word to the right or left of the search term) and “thinning” the results (i.e. the removal of irrelevant instances). In this example, the instance number 8 is irrelevant if the researcher is interested in the adjective stark; this line can then be removed from the concordance (see below for a step-to-step introduction to WordSmith). Corpus software can also usually search for words or phrases occurring within a certain distance of each other and also usually allows the use of wildcards in searches, i.e. the use of certain symbols which can represent any one character or word (if “^” is a wildcard representing one letter, a search for s'ng, for example, will yield instances of sing, sang, sung and song). The advantage of the analysis of texts with corpus software is apparent: in a few seconds the researcher receives information (about the occurrence of linguistic items in a large amount of text) that would take hours or even days if it had to be retrieved manually. The concordance-line output allows the researcher to see the occurrences in context, so that the use of the linguistic item in question, in particular frequent patterns, can often be investigated with little effort.

Retrieving more context than shown in the concordance lines, searching only a part of a given corpus, and saving the results are also regular features of almost all corpus software. Most programs can also find words frequently occurring in the vicinity of the search term and provide a list of all the words occurring in the corpus and their frequencies.

A brief introduction to an online search facility (BNC)

As mentioned above, the BNC (British National Corpus) can be accessed via the world wide web, by an online search facility allowing simple searches (available at http://sara.natcorp.ox.ac.uk/lookup.html). The search interface looks as follows:

**Simple Search of BNC-World**

Please enter your query: [input field] Solve it! Cancel

In the space provided, you can enter either words or phrases. A few more complex types of searches, such as for two words with another – unspecified – word intervening, can also be performed (a concise explanation of how to perform such more complex queries is provided on the site). Of the results, 50 at most are displayed; if there are more occurrences in the corpus, the overall number of occurrences is stated, and 50 randomly selected instances are displayed. The results are given in the context of the sentence in which they occur. If you enter the phrase not too distant, for example, the results are presented as follows:
Results of your search

Your query was
not too distant

Only 43 solutions found for this query

ADK 534 In such a climate it seems even more suitable to pursue the creation of a radical, insider ethnography and in the not too distant past this even seemed to have official support.

AB5 388 A Ferranti spokesman would not give details of the individuals or organisations against which legal action may be taken but said: ‘We would hope in the not too distant future to take the necessary legal action to put everything right.’

AJX 639 It turned out there were large surpluses in the BTR pension fund and so a good chance that the transfer value, as well as the potential pension, would be increased in the not too distant future.

APX 219 This is an interesting development and if the trend continues, we may see the demise of collateral warranties in the not too distant future.

B03 1956 ‘There must be risks that there could be moves afoot in the not too distant future for the closure of the Petersfield Court…’

If you click on the 3-digit combinations of letters and numbers given on the left, information on the text in which the sentence occurs is given.

The disadvantage of this kind of presentation is that, as the search terms are neither highlighted or aligned, patterns are not easily visible. Unlike the software that can be used with a local version of the BNC, basic functions such as sorting or thinning the results or getting more context cannot be performed, and if there are more than 50 hits not all of them can be inspected. The facility (as well as most other online corpus search facilities available) is therefore not suited for most kinds of more comprehensive linguistic research. It is, however, useful for some first practise in corpus linguistics, and for investigating some kinds of (mostly “small”) linguistic questions such as those which L2 users ask themselves when composing a text (in this case, the question might have been: Can not too distant only be used in the phrase in the not too distant future, or are there other possible combinations, such as not too distant past?). For corpus linguistic studies, this facility as well as others of its kind are useful for a first exploration of whether a certain linguistic feature is worth investigating and which questions as to the use of this feature might yield interesting results.

A step-to-step introduction to WordSmith Tools

1) Getting started

To start WordSmith (for example on the right corpus computer in the library of the English Department):

- either doubleclick on the yellow “WordSmith Tools” Icon
- or choose “Start” – “Programme” – “WordSmith Tools”

Several windows will pop up.
Your next step depends on what you want to do. The two most important options, the WordList function and the Concord function will be introduced here.

(N.B. While almost all corpora can be accessed directly through WordSmith Tools, if you want to work with ICLE (The International Corpus of Learner English), you first have to start the ICLE-program, create your own subcorpus and store it before you can access it through WordSmith; see the ICLE manual)

2) The WordList function

In the left window (“WordSmith Tools Controller”), click on “Tools”. A menu will appear; click on “WordList”. The “WordList” window will appear (which you can now enlarge to full screen size).

Now you have to select the corpus you want to work with. Let us say you want to work with FLOB (a one-million word corpus of written British English from the 1990s). Click on “File” and select “Start” from the menu. A window will pop up, where you select the option “Choose Texts Now”. Another window will pop up, displaying the structure of the files on the computer. Choose the appropriate corpus file. On the corpus computer in the library, the path for FLOB is C:\ICAME\texts\FLOB. When you have clicked your way through to your corpus (FLOB in this case), on the right of the window a list of the text files of which the corpus consists will appear (as the manual will tell you, FLOB – and all related corpora – consists of a number of subcategories, marked by a letter, such as j for science and d for religion, and all the texts belonging to one of these subcategories are stored in one file):

Click on “All” (to select all text categories included in the FLOB corpus, i.e. the whole corpus), and then on “ok”. The “Getting Started” window appears again; this time click on the option “Make a word list now”. Three new windows now pop up, on top of each other: “new wordlist (S)”, “new wordlist (A)”, “new wordlist (F)”: 
- “new wordlist (S)” gives you frequency information on each of the categories in FLOB (as well as on the whole corpus), most importantly the number of words in each category, but also a great deal of other information, such as the average word and sentence length in each category, the number of words of a certain length etc.
- “new wordlist (A)” gives you an alphabetical list of all the words occurring in the corpus, and the number of occurrences of each single word
- “new wordlist (F)” gives you the same list, only this time sorted as to frequency, with the most frequent words appearing first.

3) The Concord function
The concord function is the most important one for most kinds of linguistic analysis.

After performing the steps described in “Getting Started” above, in the left window (“WordSmith Tools Controller”), click on “Tools”. A menu will appear; click on “Concord”. The “Concord” window will appear (which you can now enlarge to full screen size).
The selection of the corpus files you want to work with works exactly as described for the WordList function above. When the corpus selection has been completed (by clicking on “All” and then “ok” as described above), the “Getting Started” window appears. Click on “specify search word”, and the following screen (“Concordance Settings”) appears:

You can now enter the word or phrase you would like to investigate under “Search Word or Phrase”. If you then click on “Go Now!”, the concordance lines of your search string appear (as illustrated for stark in the Australian Corpus of English in one of the above sections). The searches can be more complex than just for simple words or phrases, however. You can, for example, investigate the occurrence of two words within a certain span (or distance). In that case, you must enter one word under “Search Word or Phrase”, the other under “Content Word(s)”, and in addition indicate the “Search Horizons”, i.e. how many words to the left or right the second word may occur with respect the first. For example, if you are interested in the co-occurrence of the words make and decision in a span of 5 words, you enter make in the
top box, *decision* in the bottom box (or vice versa), and click on “5L” and “5R” (which tells the program to search for *decision* when it occurs either to the left or right of *make* within 5 words). If you run this search by clicking on “Go Now!” you get the following results:

If you are interested in the collocation *make a decision*, you are, however, probably also interested in clauses such as *the decision was made* etc., so that you would like to include all inflectional forms of the verb *make* (*make*, *makes*, *made*, *making*) in your search as well as all forms of *decision* (i.e. including *decisions* and perhaps *decision’s*). Rather than query the corpus for all combinations of all these forms individually, you can use an asterisk (i.e. the “*” symbol) and ask the program to search for all words beginning with the string *decision* and for all those beginning with the string *ma*. In order to do this, you have to go back to the “Concordance Settings” window by selecting “File” and “Start” (or alternatively, by clicking on the green button in the upper left hand corner), then click on “change search word”, enter *decision* in one of the boxes, and *ma* in the other, and click on “Go Now!”. This search yields 58 hits; below are the first ones (you can ignore the combinations of letters and numbers that sometimes appear in the middle of concordance lines, such as B26 210 in the second line here; they have to do with the format of the corpus, which does not concern us here):

<table>
<thead>
<tr>
<th>Concordance</th>
<th>t</th>
<th>No</th>
<th>File</th>
</tr>
</thead>
<tbody>
<tr>
<td>HAMPShIRES'S decision to make David Gower captain for the A</td>
<td>376</td>
<td>b/flob_a.txt</td>
<td>17</td>
</tr>
<tr>
<td>is particularly disappointing to have to make this A38 21 decision when ou</td>
<td>352</td>
<td>b/flob_a.txt</td>
<td>84</td>
</tr>
<tr>
<td>testing J13 47 until the child could make its own decision, in one case th</td>
<td>088</td>
<td>b/flob_j.txt</td>
<td>16</td>
</tr>
<tr>
<td>which could bring G55 163 Susan to make a decision against the relations</td>
<td>445</td>
<td>b/flob_g.txt</td>
<td>71</td>
</tr>
<tr>
<td>decisive phase - a phase when I had to make a conscious L18 70 decision</td>
<td>869</td>
<td>b/flob_l.txt</td>
<td>73</td>
</tr>
<tr>
<td>28 58 late, taking such a long time to make even the smallest decision, K</td>
<td>777</td>
<td>b/flob_k.txt</td>
<td>94</td>
</tr>
<tr>
<td>P19 the decision they made would make or break their future P19 6 to</td>
<td>159</td>
<td>b/flob_p.txt</td>
<td>61</td>
</tr>
<tr>
<td>of many of these cases in order to make an J48 138 informed decision</td>
<td>030</td>
<td>b/flob_j.txt</td>
<td>60</td>
</tr>
</tbody>
</table>
Clearly, a number of these hits are irrelevant if you are interested in the collocation *make a decision*. First of all, some words beginning with the string *ma* are not instances of the verb *make* (e.g. *may*, *matter*, *Major*). You would also probably like to exclude the noun (or adjective) *decision-making* and the noun *decision-makers*. In an instance such as number 9 (*HAMPShIRE’S decision to make David Gower captain*), *make* does not belong to *decision* syntactically; such instances also have to be excluded.

The thinning of instances (i.e. removal of irrelevant instances) works as follows: First, you click on a line showing an irrelevant instance so that it is highlighted and press the delete-button (“entf” on most German keyboards). The line thus treated becomes a lighter shade of grey (If you do that to a line by accident or discover later that you do not want a certain line removed after all, you simply have to press “einfg” to get the line back to the original colour). Once you have done this for all irrelevant instances, you click on the “zap”-button (the one that looks like a yellow flash; 7th from the right in the bottom row). The deleted instances then disappear, leaving you with 25 instances in our example:
For easier analysis, you would now like to sort the instances, so that all instances of the same form of *to make* are listed together. Sorting is done as follows: You click the “re-sort” button (4 purple arrows arranged in a circle, the third button from the left in the bottom row), and the following (“Concordance Sort”) window appears:

When you click on “Centre” in the category “Main Sort”, your results will be sorted according to the form of *make* (i.e. the central word). You can also sort the results according to the words occurring to the left or right of the search term, sort them on several levels (for
example first according to the search term and then according to the word to the right of it), and sort them according to a number of other criteria, most importantly the file names.

This screen shows the results sorted first according to the central word (“Center” in “Main Sort”) and secondly according to the words on the left of make (“1L” selected under “then by”, then “ok”).

If you want to look at more context, you click on the line in question and then click on the pink “grow” button, at the very right in the bottom row. (To get back to the original concordance lines, you can later click on the pink “shrink” button right next to it.)

In order to save the results for later analysis, click on “File”, and then “Save As”, select a drive and a folder on your computer and click on “ok”.

A (sizeable) manual explaining all the functions of WordSmith in detail is available in the library next to the corpus computers (It can also be downloaded from http://khnt.hit.uib.no/icame/manuals/wsmanual.pdf). A help function is also directly available in the program, and can be accessed from the window “Word Smith Tools Help” (which pops up in front of all others when WordSmith is started).
3) Exercises (I and II)

For step-by-step solutions click here (password required).
To obtain the password, please write an email to Nadja.Nesselhauf@urz.uni-heidelberg.de.

I Using the WordList function of WordSmith

1) What are the 5 most frequent words in Frown?

2) How frequent is the word ingredient in Frown? Is the singular or the plural form more frequent?

3) The file FROWN_P.txt contains excerpts from the category “Romance and love stories”. How many words (“tokens”) does this category consist of in the corpus?

4) Compare the 5 most frequent words in Frown to the 5 most frequent words in FLOB. N.B. When you select new texts / a new corpus to work with in WordSmith (“File” – “Start”, then “Change Selection” in the “Getting Started” window), you first have to clear your previous selection by clicking on “clear previous” in the “Choose text(s)” window! Otherwise, the new selection of texts is added to the old selection.

5) If x is the year in which you were born, what is the xth frequent word in FLOB and how frequent is it? (For example, if you are born in 1982, which is the word listed under number 1982 in “new wordlist (F)”)? Please also state what x is.

II Using the Concord function of WordSmith

1) Does the word gamut occur in Frown? If so, what are the words occurring to the right of gamut?

2) Select a word that starts with the same letter as your first name and that you think is quite frequent in English.
   a) How frequently does it occur in Frown?
   b) What is the word occurring most frequently to the left of the word (use the sorting function) and how often does it occur?

3) The word conjugal appears once in Frown. Give the full sentence in which it occurs (The sentence starts with the word In; ignore the 6-digit combinations of letters and numbers interspersed in the text).

4) What is the adjective occurring most frequently directly to the left of the word factor (use the sorting function) and how frequently does it occur?

5a) What is the easiest way of finding out whether and how often the collocation take a trip (including instances such as When we were taking that trip…, The trips we are going to take… etc.) occurs in Frown (select a distance of 6 words to both sides)?
   b) How many results does your search / do your searches yield?
c) Are all of them instances of the collocation *take a trip*? If not, give one example of a result that is not an instance of this collocation. Remove this instance (i.e. the whole line) from the concordance.
4) How to conduct linguistic analyses on the basis of corpora: two examples

Example 1: Australian English vocabulary

Let us imagine you are interested in whether Australian English vocabulary tends to be more similar to British or to American English. First, you formulate some more precise questions or hypotheses that you would like to investigate:
- Does Australian English exclusively use vocabulary items from either British or American English or does it use items from both varieties?
- If it uses items from both varieties, does it usually make a choice for a given pair of items or does it in addition mix items denoting the same concept from both varieties?
- If items from both varieties are used, can a dominance of either British English or American English vocabulary be observed?
- If Australian English uses terms from both varieties denoting the same concept in British and American English, can a difference in usage of these two items be observed?

In a next step you consult a reference book, for example Crystal’s Encyclopedia of the English Language, for differences of vocabulary in British and American English (1995: 309) and randomly choose 5 pairs of items from among those which are said to be used exclusively in one of the two varieties: flashlight (AmE) – torch (BrE), freeway (AmE) – motorway (BrE), sidewalk (AmE) – pavement (BrE), elevator (AmE) – lift (BrE), diaper (AmE) – nappy (BrE).

Then you choose your corpus. For Australian English, the Australian Corpus of English (ACE) is available to you. This corpus contains written English of different text types (you need to keep in mind, therefore, that your study has two major caveats: first, your investigation is limited to only a very small part of the Australian English vocabulary, and secondly, your investigation is based on written language only). The software you choose for the investigation is WordSmith; you start by searching for flashlight* and torch* (the asterisk is necessary to include the plural forms), and get the following results:

The next step is to investigate whether the occurrences are really occurrences of the concept in question. In the case of flashlight, the co-occurrence of the words with pitch darkness indicate that some kind of lamp is indeed referred to. In the case of torch, the first two occurrences appear to be part of a proper name (that refers to some kind of operation) and therefore need to be disregarded. The third instance that has been thrown up by the search is neither an instance of torch nor of torches; the fourth is an instance of the verb to torch. These two instances need to be disregarded. The final two instances actually are instances that seem to refer to a portable lamp. When you take a closer look at the files in which the instances
occur (under “File” on the right of the window), it turns out that \textit{torch} occurs in two different categories, a and k (and therefore in two different texts), whereas the two instances of \textit{flashlight} occur in the same category (category b). If you look at the context (you might want to “grow” the line to receive more context), it turns out that they occur in the same text, in two consecutive sentences. A common notation for this is $2 \ (1)$, i.e. two occurrences in one text.

You then perform searches in the same fashion for the other words: \textit{freeway*}, \textit{motorway*}, \textit{pavement*}, \textit{sidewalk*}, \textit{elevator*}, \textit{lift*}, \textit{diaper*}, \textit{napp*} (N.B. a search for \textit{nappy*} would not throw up the plural form in this case, so you either have to perform two searches, for \textit{nappy} and \textit{nappies}, or a search for \textit{napp*}) These searches give you the following output:

A close inspection of the contexts of these instances leads you to the conclusion that they are all instances of the relevant concept.

There are no results for \textit{motorway*}.

All instances of \textit{sidewalk*} and \textit{pavement*} appear to refer to the relevant concept (again, if you are unsure about certain instances you can retrieve more context to decide).

There are no results for \textit{elevator*}.

For \textit{lift*}, you get 66 results. Glancing through them, you notice that many of these are instances of the verb \textit{lift} rather than of the noun. One easy way to get rid of a number of the verb-instances is to perform the search again, only this time by excluding the form \textit{lifted} (in the window “Concordance Settings”, type in \textit{lifted} in the “but excluding”-box, which appears whenever you type in a string containing an asterisk in the “Search Word or Phrase” box). This leaves you with 40 instances. The form \textit{lifting} is not relevant to your investigation either,
so you sort the instances according to the central word, and remove all instances of lifting. As you glance through the remaining instances, you notice that many instances of the verb lift are preceded by to. You perform another sort, this time according to the word to the left of the search term, and remove the instances of lift preceded by to, which leaves you with the following 24 instances:

You go through these one by one and decide whether they are instances of the relevant concept. While doing that, you notice that it is not only the verb to lift that has to be excluded from the count, but also other meanings of the noun lift, in particular the meaning as in to give sb. a lift and those instances referring to a ski lift (if you are uncertain whether there is a British-American difference for this term, consult a dictionary). To decide whether certain instances refer to a ski lift or a lift in a building, you will need to look at more context in some cases. In the end, you are left with 6 instances (numbers 1, 2, 4, 18, 19, 20).

There are no results for diaper*.

All 5 instances of nappy appear to be instances of the concept in question.

To get an overview over your results, you enter them in a table.

<table>
<thead>
<tr>
<th>words:</th>
<th>occurrences of British word:</th>
<th>occurrences of American word:</th>
</tr>
</thead>
<tbody>
<tr>
<td>torch : flashlight</td>
<td>2</td>
<td>2 (1)</td>
</tr>
<tr>
<td>motorway : freeway</td>
<td>0</td>
<td>7</td>
</tr>
</tbody>
</table>
As with the first example (flashlight and torch), you could check whether in those cases with several occurrences, the words occur in the same text or not. As all the texts of one category are stored in one file in this corpus, this is difficult to find out in some cases, however. If the instances occur in different categories, they also appear in different texts. If they appear in the same category, a good indication is the word number (which is given next to the file name in the concordance line; a word number such as 5,212, for example, means that in the category in question the search term is the 5,212th word). As the texts in the Australian Corpus of English are only around 2,000 words long (see the corpus manual), two instances that have word numbers more than 2,000 words apart are unlikely to occur in the same text. If they have word numbers that are apart less than 2,000 words, one way to find out is to open the file with a word processor and use the search function of that program. For the purposes of this investigation, you judge this as unnecessary, however.

What do these results tell you with respect to your research questions then?

First, Australian English does not exclusively use either British or American vocabulary. Secondly, the results indicate that in some cases, in Australian English either the British or the American word is used exclusively (but the results are no conclusive proof: a larger corpus might throw up instances of the other variety): In the case of freeway – motorway, only the American word occurs in the corpus, in the case of nappy – diaper and lift – elevator, only the British word occurs. This might lead to the hypothesis (which in turn might be the basis for further research) that words relating to traffic and cars might be dominantly American in Australian English. In some cases, however, Australian English uses both the American and the British term. In the case of torch and flashlight, the words might be used with similar frequencies (though the numbers are too small for any firm conclusions). In the case of pavement and sidewalk, pavement appears to be the dominant variant. Moreover, it appears from the results that sidewalk is only used in modifying function (sidewalk throng, sidewalk stalls, sidewalk cafes and bazaars). None of the instances of pavement occurs in this function. So this might be a case where the words from the two varieties have undergone a functional distinction in Australian English.

Although the words investigated are of course too few for any definite conclusions, the results (see the totals in the table) indicate that while British vocabulary is dominant in Australian English, there has also been some influence of American English. (In a larger study, you would of course not only include more items, but also contrast your results with those recorded in relevant literature on Australian English).
Example 2: Present perfect and simple past in British and American English

You are interested in differences of the use of the present perfect and the simple past in British and American English. In Swan, *Practical Guide to English Usage*, you find that in American English the simple past can be used with adverbs such as *just*, *already*, *yet*, *ever*, and *before*, whereas in British English only the present perfect is possible (21995: 423). First, you formulate some questions or hypotheses on what precisely you want to find out:

- Does the simple past not occur at all with these adverbs in British English?
- Does American English allow only the simple past, or also the present perfect?
- If American English allows both the simple past and the present perfect, which is more frequent (and how much more frequent exactly)?
- If American English allows both, when is what tense / aspect used (i.e. does the choice depend, for example, on the verb of the clause in question, are there slight differences of meaning depending on which tense / aspect is used etc.)?

Then you choose corpora on which to base your analyses. Since you are interested in Present Day English, and in American and British English, two comparable corpora of the two varieties in question would be ideal. From the list provided on this website you glean that the two most suitable corpora for your investigation are FLOB and Frown. FLOB and Frown are two matching corpora of British and American English of the 1990s, which contain different written text types. (This is one important caveat of your study, then, which should be kept in mind: you are only investigating written English and not spoken English). The software you chose for your investigation is WordSmith. You decide to start your investigation with the adverb *just*. In order to answer the first three questions, you first look at the relative occurrence of the simple present and the present perfect with *just* in both varieties. First you search for *just* in FLOB and Frown: In FLOB you find 1107 occurrences, in Frown 1124. Too many to look at individually. What can you do?

One option is to choose a few verbs which are frequent and have the program look up occurrences of these verbs which occur next to *just*. You choose the verbs *be*, *do*, *go*, *give*, *hear*, and *say*. The next point you have to consider is where in relation to *just* and in which form the verb usually occurs in a clause when *just* refers to the past. *Lucy has just called* is the example given in Swan (1995: 419) for British English; *Lucy just called* the example for American English. So the form of the verb occurs to the right of *just* (*Lucy called just* or *Lucy just has called* or *Just Lucy has called* are very unlikely to occur with the ‘very recently’ meaning of *just*). As the present perfect is formed with the past participle and the simple past with the past tense form (which are not always identical), and as the past tense has two forms for the verb to *be*, the following forms will have to be investigated: *been, was, were; done, did; gone, went; given, gave; heard; said*. You first investigate the usage in FLOB (the British corpus).

*Just been* yields 10 instances:
Just was and just were yield 0 instances.

Just gone yields 2 instances:

Just went yields 0 instances.

Just done yields 1 instance:

Just did yields 0 instances.

Just given yields 1 instance:

Just gave yields 0 instances.

Just heard yields 0 instances.

Just said yields 4 instances:

The next step is a careful consideration of the occurrences. First of all, the instances of the past perfect (had just + past participle) have to be excluded, such as had / 'd just been. In one case, Perhaps he just been unlucky, it cannot be decided whether this is an instance of past perfect or present perfect, so it has to be disregarded. One instance, having just been abandoned is a participle clause, which has to be disregarded as well (as participle clauses are
non-finite and therefore tenseless). Then, the semantics of just have to be considered. In two cases at least (there’s just been the two of us and I just said that to tease Oster), just clearly means ‘only’ and not ‘very recently’, so these instances have to be disregarded as well. For the instances where the concordance lines do not provide enough context to exclude the ‘only’-meaning of just with some definiteness, you have to look at more context to decide. The instance has just gone bust, for example, could be both an instance of the ‘only’-meaning and of the ‘very recently’-meaning of just, but the context seems to support a ‘very-recently’-reading. In I must have just gone, however, just turns out to be an instance of the ‘only’-meaning.

The manual sifting thus leaves you with the following results for FLOB:

<table>
<thead>
<tr>
<th>forms:</th>
<th>number of occ.</th>
</tr>
</thead>
<tbody>
<tr>
<td>just been</td>
<td>5</td>
</tr>
<tr>
<td>just was / were</td>
<td>0</td>
</tr>
<tr>
<td>just done</td>
<td>0</td>
</tr>
<tr>
<td>just did</td>
<td>0</td>
</tr>
<tr>
<td>just gone</td>
<td>2</td>
</tr>
<tr>
<td>just went</td>
<td>0</td>
</tr>
<tr>
<td>just given</td>
<td>1</td>
</tr>
<tr>
<td>just gave</td>
<td>0</td>
</tr>
<tr>
<td>just heard (present perfect)</td>
<td>0</td>
</tr>
<tr>
<td>just heard (simple past)</td>
<td>0</td>
</tr>
<tr>
<td>just said (present perfect)</td>
<td>0</td>
</tr>
<tr>
<td>just said (simple past)</td>
<td>1</td>
</tr>
</tbody>
</table>

In FLOB, therefore, of the 9 instances of the 6 verbs with the adverb just, 8 take the present perfect, and one the simple past. When we look at the instance in the simple past (I thought you just said that she was beautiful) more closely, however, it turns out that it is most likely a case of a backshifted present. As our numbers are small, we cannot, therefore, say with confidence whether the simple past tense does not occur with just in British English (except as backshifted present) or whether it is just much less frequent than the present perfect. What we can say is that our results indicate that the present perfect is clearly predominant in (written) British English.

You then perform the same searches in Frown, with the following output:

**Just been:**

<table>
<thead>
<tr>
<th>Concordance</th>
<th>Set</th>
<th>d No.</th>
<th>File</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>tonal beauty_one who had G36</td>
<td>9.1</td>
<td>just been on safari in Africa with 14</td>
<td>87.5</td>
<td>frown_g.txt 48</td>
</tr>
<tr>
<td>Professor Byron 40</td>
<td>70</td>
<td>Snipes had just been published in the avant-garde</td>
<td>8.2</td>
<td>frown_m.txt 57</td>
</tr>
<tr>
<td>28 Warner Brother's cartoon that had just been smashed by a flying N28</td>
<td>2</td>
<td>2</td>
<td>67.1</td>
<td>frown_n.txt 93</td>
</tr>
<tr>
<td>particularly in light</td>
<td>K15</td>
<td>203 of what I've just been remembering about my mother</td>
<td>36.6</td>
<td>frown_k.txt 52</td>
</tr>
</tbody>
</table>

Of these, three are instances are instances of the past perfect; the last one is an instance of the present perfect.

The searches for just done and just did yield no instances.

**Just gone** is represented by one instance, which again is an instance of the past perfect, however:
There is one instance of *just went*:

This, however, is an instance of the ‘only’- meaning of *just* and therefore has to be disregarded.

*Just given* has one instance, which is, however, an instance of the past perfect:

Just gave:

Here, the instances of *just* meaning ‘only’ have to be excluded. If you look at more context, it becomes clear that only instance number 3 is an instance of the ‘very recently’ meaning.

There are two instances of *just heard*, both instances of the simple past:

*Just said* yields two instances:

While the first of these is an instance of the past perfect, the second actually seems to be one of the ‘very recently’ meaning if the context is inspected (*he just said clucking – ‘We clucking appreciate it’*).

<table>
<thead>
<tr>
<th>forms</th>
<th>number of occ.:</th>
</tr>
</thead>
<tbody>
<tr>
<td>just been</td>
<td>1</td>
</tr>
<tr>
<td>just was / were</td>
<td>0</td>
</tr>
<tr>
<td>just done</td>
<td>0</td>
</tr>
<tr>
<td>just did</td>
<td>0</td>
</tr>
</tbody>
</table>
In Frown, of the 5 occurrences of the 6 verbs 4 therefore are in the simple past and one in the present perfect. As the numbers are small, we have to be careful with our conclusions. It seems from the results, however, that in American English the simple past is indeed predominantly used with just but that the present perfect also occurs. Due to the small numbers, the fourth question asked at the beginning cannot be answered on the basis of this analysis only.

<table>
<thead>
<tr>
<th>Verb</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>just gone</td>
<td>0</td>
</tr>
<tr>
<td>just went</td>
<td>0</td>
</tr>
<tr>
<td>just given</td>
<td>0</td>
</tr>
<tr>
<td>just gave</td>
<td>1</td>
</tr>
<tr>
<td>just heard (present perfect)</td>
<td>0</td>
</tr>
<tr>
<td>just heard (simple past)</td>
<td>2</td>
</tr>
<tr>
<td>just said (present perfect)</td>
<td>0</td>
</tr>
<tr>
<td>just said (simple past)</td>
<td>1</td>
</tr>
</tbody>
</table>

**What you have to take into account when performing a corpuslinguistic analysis**

As the above analyses have shown, there are certain points that you need to take into account in each corpus-based analysis:

- It is essential that you know what exactly is in the corpus (e.g. what kinds of text, what variety, how many words a text contains, what the different files contain, what the abbreviations in the file names stand for etc.); so always have a look at the corpus manual before you start your work
- If you use different corpora for an analysis, find out to what degree they are comparable (corpus size, variety, time span covered, text types) etc.
- You need to make sure that your search (or searches) actually finds all occurrences of the language item in question (for example all inflectional variants)
- After performing a search with corpus software, you have to check whether all results are actually relevant for the investigation in question (and if not which ones have to be excluded)
- You need to check how the item in question is distributed in the corpus (e.g. whether it only occurs in certain periods / text types etc.); it is also relevant from how many different texts (or text categories) the solutions come from (i.e. if an item is very frequent, but occurs only in one or two texts, this does not mean that it is frequent in the variety / text type etc. investigated)
- If an item does not occur in a given corpus, this does not necessarily mean that it does not occur in the variety / text type / period etc. represented in the corpus, especially if the corpus is fairly small
- Small differences in frequency can be a matter of chance and therefore need to be interpreted prudently
- In the interpretation of your results, be aware of what you have actually investigated (e.g. only written or only spoken language, only certain instances of the phenomenon in question etc.)
5) Exercises (III)

Exercise III.1:

Continue the investigation started in Example 2. Choose another adverb from the list given in Swan (see above, first paragraph in Example 2) and see whether the results for this adverb can contribute further to answering the four research questions. Give not only your results but state how you have arrived at them.

Exercise III.2:

For step-by-step solutions click here (password required).

To obtain the password, please write an email to Nadja.Nesselhauf@urz.uni-heidelberg.de.

You are interested in the recent development of whom in British English. On the basis of the corpora LOB (1961, British English) and FLOB (1991, British English), you attempt to answer the following two questions:

a) Was there an increase or a decrease (or neither) of whom in the course of these 30 years?

b) Does the (possible) development affect those instances of whom after a preposition (to whom, with whom etc.) to the same degree as those without?

(N.B. LOB and FLOB contain the same text types and the same number of words, so that the results are directly comparable)

(N.B. when analysing the concordance lines, ignore the letter-number complexes that sometimes interrupt the text; these may also affect the sorting function)

Describe briefly what you have done in order to find answers to these questions and present your results in the following table as well as verbally:

<table>
<thead>
<tr>
<th></th>
<th>whom (total)</th>
<th>whom (after preposition)</th>
<th>whom (without preposition)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOB (1961)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FLOB (1991)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
6) Where to find further information on corpus linguistics

Other web-based introductions to corpus linguistics:
http://www.georgetown.edu/faculty/ballc/corpora/tutorial.html
(Concordances and Corpora Tutorial by Catherine Ball at Georgetown University.)
http://www.ling.lancs.ac.uk/monkey/ihe/linguistics/contents.htm
(By McEnery & Wilson, intended as a supplement to their 2001 book.)

General books on corpus linguistics:

Collections on studies and issues in corpus linguistics:

A few examples of (paper-length) corpus-based studies


Information on certain corpus resources:


On corpus linguistics versus introspection:

On statistics and corpus linguistics:

On corpus linguistics and linguistic theories:


On the use of corpora for diachronic analyses:

On the analysis of learner corpora: